Generating and using probabilistic morphological resources for the biomedical domain

Vincent Claveau, Ewa Kijak

Generating and using probabilistic morphological resources for the biomedical domain

Vincent Claveau[†], Ewa Kijak[*]

IRISA – [†]CNRS – [*]Univ. of Rennes 1
Campus de Beaulieu, Rennes, France
vincent.claveau@irisa.fr ewa.kijak@irisa.fr

Abstract

In most Indo-European languages, many biomedical terms are rich morphological structures composed of several constituents mainly originating from Greek or Latin. The interpretation of these compounds are keystones to access information. In this paper, we present morphological resources aiming at coping with these biomedical morphological compounds. Following previous work (Claveau and Kijak, 2011; Claveau, 2012), these resources are automatically built using Japanese terms in Kanjis as a pivot language and alignment techniques. We show how these alignment information can be used for segmenting compounds, attaching semantic interpretation to each part, proposing definitions (gloses) of the compounds... When possible, these tasks are compared with state-of-the-art tools, and the results show the interest of our automatically built probabilistic resources.

Keywords: Morpho-semantic analysis, biomedical terminology, probabilistic morphological resources

1. Introduction

In the biomedical field, specialized terms are keystones to access information. However, in most Indo-European languages, these terms are rich morphological structures composed of several constituents mainly originating from Greek or Latin. Such morphological complexity is important to take into account for basic processes (translation, establish semantic relations...) as well as higher-level HLT, like machine translation or Information Retrieval (IR).

In this paper, we present morphological resources aiming at coping with these biomedical morphological compounds. Following previous work (Claveau and Kijak, 2011; Claveau, 2012), these resources are built automatically from the UMLS MetaThesaurus (Tuttle et al., 1990) (a multilingual database grouping several biomedical terminologies), and are now available in several languages (including, English, French, Spanish...). Our approach relies on the use of Japanese as the pivot language, and more specifically on terms written in kanjis to help the decomposition of words in other languages. In a fully automatic way, they are cut into morphemes and each morpheme is associated with the corresponding kanjis.

For example, the term photochemotherapy can be translated into Japanese by 光化学法; indeed, by decomposing and aligning these two terms, we have:

- photo ↔ 光 ('light');
- chemo ↔ 化学 ('chemistry', 'drug');
- therapy ↔ 法 ('therapy') .

As we see here, each morpheme is associated with kanjis that can be seen as semantic descriptors, more suitable for NLP problems than the initial full term. Thus, our morphological analysis is chiefly based on the alignment between morphemes and kanjis which is performed with a forward-backward algorithm adapted to the manipulated data. As a side effect, this process generates a probabilistic correspondence table (for each language) between kanjis and morphemes. These tables are the main component of the morphological resources presented in this paper, as they can be used in many applications as we will see. In addition, this paper also describes how the UMLS can be used again to attach a biomedically suited translation to the morpheme/kanjis pair (eg. the morpheme hemo, aligned to 血液, is translated by 'bloody').

The remaining of the paper is structured as follows. We first review related work and resources (Sect. 2.). In Sect. 3., we describe the processes that make it possible to generate the morphological resources for most European languages. In Sect. 4., several evaluations are reported on different aspects of the generated resources.

2. Related work

Different studies rely on morphology to perform some terminological analysis. It is especially the case in the biomedical domain. Indeed, on the one hand, terminologies are a keystone for many applications, and on the other hand, those terms are usually built by a morphological operation said neoclassical composition. For example, terms like magnetoencephalography can be decomposed into three morphs: magneto/encephalo/graphy. The nature of the implied morphological units (Iacobini, 1999; Dal and Amiot, 2008) as well as the rules governing the way these units are composed (Dal and Amiot, 2008; Fradin, 2005) make these morphologically-complex terms particular linguistic objects. Beside that, the many terms that are built this way and the productive property of the neoclassical composition (it serves to produce many neologisms) make this phenomenon also important from an applicative point of view.

There exists some databases containing morphological information for the biomedical domain, like Biotop[1] in

[1] georges.dolisi.free.fr/

French or Dorland's[2] in English. Yet, these resources are far from being complete, and their use for automatic analysis is difficult (each morpheme is solely described with an informal definition) and no information is given on how to combine them.

Apart from that, morphological analysis systems have been proposed. Some of them adopt a lexematic approach, in which the term forms are used to exhibit relations between terms, but without decomposing explicitly into morphemes (Grabar and Zweigenbaum, 2002; Claveau and L'Homme, 2005; Hathout, 2009, for example). Other studies adopt a morphemic approach in which terms are decomposed into morphemes. These studies can be grouped according to the amount of expert knowledge or resources on which they are based. Some fully automatic techniques only necessitate a list of terms in which recurrent letter sequences will be considered as morphemes (Creutz and Lagus, 2005; Kurimo et al., 2010, *inter alia*). Yet, such approaches cannot associate any knowledge with the morphemes. Other morphemic work relies on expert knowledge: morphemes, their semantic information, and the morphological composition rules are mostly manually given as facts or heuristics (Baud et al., 1999). Among this family of approach, one can further distinguishes the techniques according to their applicative goal and the more or less detailed analysis that they provide. For instance, some provide a segmentation of the morphological compounds and assign interlingua identifiers to the morphological units found (Markó et al., 2005). Other techniques provide a hierarchical decomposition and an interpretation of the compound (Namer, 2007; Deléger et al., 2008). The work presented in this paper adopts the same morphemic approach, but our resources are automatically built, which makes them more complete and easily available in many languages.

From a technical point of view, the approaches used to generate and exploit our probabilistic morphological resources can be compared with existing work on transliteration, in particular of Arab or of terms written in Katakana, for both direct translation (Knight and Graehl, 1998, for example), or for the search of translation (Chiao and Zweigenbaum, 2002; Tsuji et al., 2002). From this point of view, a close work is proposed by (Morin and Daille, 2010); they propose to find equivalence between French terms and Japanese terms written in Kanjis based on morphological considerations. Yet, here again, the few rules at the heart of their approach have to be built by an expert, and only a simple morphological phenomenon (derivation) is considered. Their approach cannot be used for neo-classic compounds as we are aiming to do.

3. Generating biomedical morphological resources from the UMLS

3.1. On the use of kanjis as pivot

As explained before, the generation of our morphological resources relies on an alignment step of terms from the studied language with their translations in kanjis. It is worth noting that the choice of kanjis as pivot language is not fortuitous. Kanjis, one of the three Japanese writing systems, are very used in specialized domain; inherited from the Chinese sinograms, they are mostly pictograms or ideograms. They are suited to our problem, since they are invariable whatever their position in the term and neighboring kanjis. Obviously, they are also independent from Latin and Greek, which prevent our approach to find fortuitous regularities based on common etymology.

Our approach makes a strong hypothesis of parallelism: the kanji terms have to be built in a similar way than the morphological composition of the terms in the other language. This hypothesis is verified in most cases since the Japanese word order and the morphological composition rules at stake are identical. Indeed, Japanese is sometimes qualified as a free-order language, but one of its basic principles is to indicate the characteristic of an object before the object, arguments before the predicate, and more generally the governed before the governing (Nakamura-Delloye, 2007). This order is also the one used in Roman languages for neoclassical composition (on the contrary of ordinary composition) in which the semantically governed morphological unit is before the governing one (Dal and Amiot, 2008).

3.2. Previous work: alignment

In previous studies (Claveau and Kijak, 2011; Claveau, 2012), we have proposed an original approach to align morphologically-complex compounds with their Kanjis translations. The only requirement is a list of terms in the studied language with their Kanjis translations, without any pre-processing. The result of our alignment approach produces at the same time the decomposition of the terms into morphs and aligns the morphs to the corresponding kanjis.

This alignment is performed with an *Expectation-Maximization* (EM) algorithm that we briefly present hereafter (Jiampojamarn et al., 2007, for more details and use examples). It is based on a *Baum-Welch* algorithm (Rabiner, 1989), more precisely a forward-backward algorithm, extended to allow the alignment of sub-sequences of symbols instead of 1-1 alignments only. The principle of the algorithm is to alternate two operations: the first one computes a table of counts recording which possible alignment is seen with a weight based on the probability of this alignment in each training pair. The second step estimates the alignment probabilities based in turn on the table of counts. These two steps are detailed more formally in Algorithm 1. The first one (Expectation, in Algorithm 2) processes each pair (x^T, y^V) of term with its Kanji translation to be aligned (T and V are respectively the number of letters in the term x and number of Kanjis in the Japanese translation y) with a Forward-backward approach (Algorithm 4), and outputs an updated table of counts named γ. The second step, the Maximization (Algorithm 3), uses this table of counts to produce an alignment probability table named δ. This table is used in the remaining of the paper to propose morphological analyses; an excerpt of this table in given in Figure 1. For the interested reader, more details are available in (Claveau and Kijak, 2011; Claveau, 2012).

[2] www.dorlands.com/wsearch.jsp

Algorithme 1 *EM alignment algorithm*

Input: list of pairs (x^T, y^V), $maxX, maxY$
while changes in δ **do**
  initialize γ to 0
  for all pair (x^T, y^V) **do**
    γ = Expectation($x^T, y^V, maxX, maxY, \gamma, \delta$)
  δ = Maximization(γ)
return δ

Algorithme 2 *Expectation*

Entrée : (x^T, y^V), $maxX, maxY, \gamma$
α := Forward-many2many($x^T, y^V, maxX, maxY, \delta$)
β := Backward-many2many($x^T, y^V, maxX, maxY, \delta$)
if $\alpha_{T,V} > 0$ **then**
  for $t = 1...T$ **do**
    for $v = 1...V$ **do**
      for $i = 1...maxX$ t.q. $t - i \geq 0$ **do**
        for $j = 1...maxY$ t.q. $v - j \geq 0$ **do**
          $\gamma(x^t_{t-i+1}, y^v_{v-j+1})$ +=
          $\frac{\alpha_{t-i,v-j}\delta(x^t_{t-i+1}, y^v_{v-j+1})\beta_{t,v}}{\alpha_{T,V}}$
return γ

Algorithme 3 *Maximization*

Input: γ
for all sub-sequence a s.t. $\gamma(a, \cdot) > 0$ **do**
  for all sub-sequence b s.t. $\gamma(a, b) > 0$ **do**
    $\delta(a, b) = \frac{\gamma(a,b)}{\Sigma_x \gamma(a,x)}$
return δ

Algorithme 4 *Forward-many2many*

Input: (x^T, y^V), $maxX, maxY, \delta$
$\alpha_{0,0} := 1$
for $t = 0...T$ **do**
  for $v = 0...V$ **do**
    if $(t > 0 \vee v > 0)$ **then**
      $\alpha_{t,v} = 0$
    if $(v > 0 \wedge t > 0)$ **then**
      for $i = 1...maxX$ s.t. $t - i \geq 0$ **do**
        for $j = 1...maxY$ s.t. $v - j \geq 0$ **do**
          $\alpha_{t,v}$ +=
          $\delta(x^t_{t-i+1}, y^v_{v-j+1})\alpha_{t-i,v-j}$
return α

3.3. Probabilistic morphological information

The probabilistic correspondence tables between morphemes and kanjis are generated by an alignment process based on the forward-backward algorithm presented in the previous section. Beside the alignment results, it produces the δ table (see the excerpt given in Figure 1), containing the alignment probability iteratively computed from the pairs of terms associated with their Kanji translations. Thus, in practice, the algorithm only needs a list of pairs of terms (of the studied language) with their Kanji translations. Such lists are easily extracted from the UMLS as each term is associated with a CUI (*Concept Unique Identifier*), which is independent of the language. When building

```
上/ia;              0.00099
上嫌/euphor         4.950495e-05
上嫌/euphoria;      4.950495e-05
上炎/itis;          4.470132e-52
上狭窄/ostenosis;   5.59957e-23
上狭窄/stenosis;    7.716783e-17
上皮/carcino        2.568568e-311
...
```

Figure 1: Abstract of the probability table δ for English produced by the alignment algorithm.

these lists, we focus on single-word terms which are more likely morphological compounds than multi-terms.
Another probabilistic information is given by our alignment algorithm: it is possible to study how often a particular morph is aligned with a particular sequence of Kanjis. The more they are aligned with each other, the more semantically close they may be considered. All these links form a graph in which the nodes are the Kanjis or the morphs, and the edges are weighted according to the alignment frequencies provided by the alignment algorithm on the UMLS pairs. Figure 2 shows a small excerpt of this graph; the thickness of the links is proportional to edge weight.

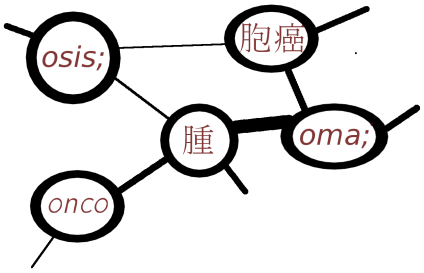

Figure 2: Morphosemantic graph of morphs-kanjis

We have shown in previous work how this graph could be mined to discover semantic relations between the morphs (Claveau, 2012). In particular, first order relations are morphs that are close to each other in this weighted graph. In this paper, we are also interested in second-order relations: morphs sharing the same (first-order) neighbors. An example of these second-order neighbors for the French morph gastro (stomach) is given in the form of word cloud in Figure 3.

3.4. Attaching relevant translations to kanjis

One limit of the probability tables, as produced by the algorithm explained above, is that the kanjis are not understandable by non Japanese speakers. This does not prevent their use in applications like information retrieval (Claveau, 2012), but they cannot be used for providing human-readable interpretation of the morphological compounds. It is of course possible to use bilingual dictionaries, but they may not provide the most relevant translation given our biomedical domain (for instance, 路, commonly trans-

Figure 3: Cloud of the second-order neighbors of the French morph gastro

	炎	any other kanjis
inflammation	a	b
any other word	c	d

Table 1: Cooccurrence values to compute the strength of the association 炎/inflammation

lated by 'road' or 'way', is better translated in the biomedical context by 'tract'). We propose instead to use again the UMLS to generate relevant translations. We exploit again the CUI to collect pairs of terms with their translations, but here, only multi-word terms are considered (see Fig. 4). Indeed, these terms are usually composed of morphologically simpler and more common words than the morphological compounds.

oropharyngeal scar :: 口腔咽頭瘢痕
infection; navel, newborn :: 新生児臍炎
reproductive and/or childbirth services :: 助産診断学
biopsy heart normal :: 心臓生検正
career choice :: 職業選択
increased diastolic arterial pressure :: 拡張期血圧高値
foreign body in the lacrimal punctum :: 涙点内異物
...

Figure 4: Abstract of the multi-word English terms/kanjis list extracted from the UMLS.

To associate these simpler words with the kanjis, one could use again an alignment process, but it is made more difficult since the hypothesis of parallelism does not hold here. Instead, we use a simple method based on the cooccurrences of words and (any substring of) kanjis in any pair of the list. As it is done for collocation extraction, many statistical indices based on these cooccurrence counts (Pearce, 2002) can be used to measure the strength of the relation between each word and each kanji combination. In this work, we use the point-wise Mutual Information (MI) score which has proved useful for many applications. Consider the example given in Tab. 1, the MI score for 炎 / inflammation is $MI = \log \frac{a}{(a+c)*(a+b)}$. The retained translation for a given kanji or combination of kanjis is the word that maximizes its MI score.

4. Experimental validation

As the resources are automatically built, it is important to evaluate their relevance in real tasks. In this paper, we present evaluation through three tasks (linear segmentation of terms, hierarchical structuration of terms, morphological analysis).

4.1. Linear segmentation of the morphological compounds

Segmenting a morphological compound into morphemes can be useful for many applications. Our probabilistic resources can be used for such a task with a simple Viterbi algorithm which will propose the most probable decomposition of the term, based on the probabilities of each element found in the table. In order to evaluate the precision of this approach, a ground-truth has been developed for French terms; it is composed of 1 000 terms that have been manually segmented into morphemes. The Viterbi algorithm is thus given the probability table generated by the alignment algorithm applied to French/Kanji pairs extracted from the UMLS. As baselines, we also compare this approach with two well known tools of the domain: Morfessor[3] (Creutz and Lagus, 2005) and DeriF[4] (Namer, 2007). Morfessor being based on a learning approach, it is given the list of (non multi-word) French terms extracted from the UMLS (about 13 000 terms). As illustrated in Fig. 5, our approach, based on automatically generated resources, rivals the expert-based system DeriF, and outperforms MorphoSaurus.

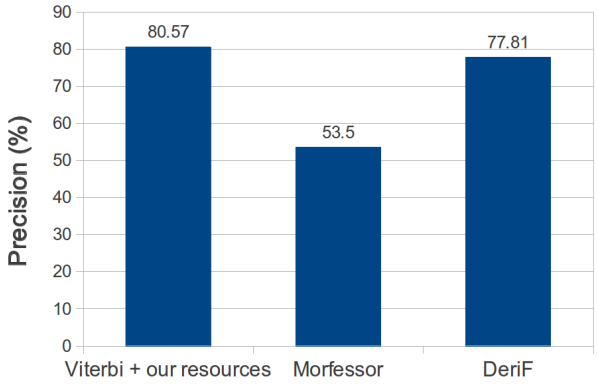

Figure 5: Precision (%) of term segmentation on French.

It is also interesting to examine how the precision of the systems evolves with respect to the amount of training data used, that is, the number of words for Morfessor, and number of pairs for our system. These results are presented in Figure 6. The results of Morfessor are almost constant whatever the amount of training words processed. This somewhat surprising result is explained when examining the errors: Morfessor tends to over-segment on the basis of fortuitous similarity between words. When given the same amount of training data, our approach yields better results, but it should be noticed that it also exploits more information though the kanji translations. DeriF's results are close

[3] www.cis.hut.fi/projects/morpho
[4] www.cnrtl.fr/outils/DeriF/requete.php

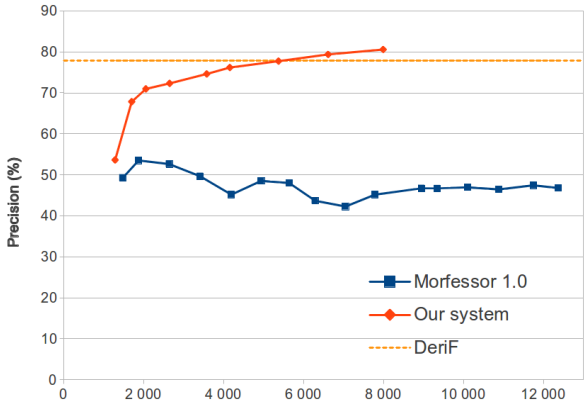

Figure 6: Precision (%) on the segmentation task on French terms according to the amount of training data (number of terms for Morfessor, number of Kanji-French pairs for our approach).

to ours on this segmentation task. Most of its errors are absence of segmentation; it may be explained by the presence of morphs that are unknown to DeriF (*e.g.* améloblastome, athétose, micrognathisme, sparganose, sérosite are not segmented), or by certain configurations of morphs (argyrose, cholangite, angiocholite). In rarer cases, the proposed segmentation is wrong; this is due to incorrectly recognized morphs (*e.g.* ré/tinoblastome). The absence of segmentation, or under-segmentation are also the major causes of errors of our approach. In our case, they are due to the presence of rare morphs or kanjis that are seldom represented in the training data. Obviously, these errors tend to diminish when the training dataset is larger. On the whole, DeriF and our approach agree on 70,5% of the segmentation, and more than 80% of the correct segmentations found by one are also found by the other.

4.2. Structural (hierarchical) analysis

The segmentation analyses examined in the previous subsection are linear: the morphological units are put on a same level. Yet, the morphological compounds have an inner hierarchical structure which brings important information to understand the term. As it was mentioned in Section 3.1., the order used to build the biomedical morphological compounds, as other scientific compounds, is particular in Roman languages. In general, the governed morphological units appear before (i.e. at the left of) their governing units (Dal and Amiot, 2008). It seems reasonable to use this general rule as a default structuring rule; it is noted rule 1 hereafter. Each morph is thus considered as a modifier of the compounds it precedes, which may be written as follows:

$$m_1 m_2 ... m_n = [\ m_1\ [\ m_2\ [\ ...\ [\ m_n\ ]\ ]\ ]\ ]$$

Of course, there are exceptions. One of the most common is the structure of terms like meningoencephalitis whose structure should be:
[[[meningo][encephal]] itis]
or adenocystoma that is structured as:
[[[adeno][cysto]] oma].
To decide whether this structuring rule should be used instead of rule 1, we simply look at the morph appearing in the term. If two consecutive morphs are found to be second-order neighbors (i.e. at least one of them is in the 10 closest neighbors of the other; see Sec. 3.3.), they are considered in parallel, that is like meningo and encephal, and not in the governing-governed relation of rule 1. This modification to rule 1 is noted rule 2.
Even with this second structuring rule, there are other exceptions such as angiocardiography, whose structure should be:
[[angio [cardio]] graphy].
Such terms would thus require additional rules, but in the experiment reported hereafter, only the two previous rules were implemented.
To evaluate the relevance of this structuring approach, we have conducted a small experiment on about 200 French terms that were manually structured by the authors. In order to evaluate the quality of the structuring rules only, our approach was given the correct segmentation of the terms. The performance is evaluated in term of precision, that is the amount of terms completely and correctly structured by this approach. As a comparison, we also indicate the results obtained by DeriF; for a fair comparison, only terms that were correctly segmented by DeriF are considered in these structuration results.

Method	rule 1 only	rule 1 + rule 2	DeriF
Precision	63.1	76.2	78.4

Table 2: Precision (%) of different methods to hierarchically structure biomedical morphological compounds.

These results highlight the preponderance of the governed-governing scheme (denoted rule 1) in the construction of the neo-classical morphological compounds. With the addition of the rule 2, that relies on the second-order affinities, we rival the performance of expert systems like DeriF which heavily rely on expert knowledge.

4.3. Morphological analysis

Beside the segmentation performance, it is also interesting to investigate the semantic interpretation given by the kanjis and their translations (generated as explained in Sect. 3.4.). Building a direct evaluation is difficult but comparing the analysis with DeriF is insightful. We show in Tab. 4 some analyses provided by DeriF that may benefit from our more complete resources (the DeriF gloses, initially in French, are translated by us, words in italics are errors or invention by DeriF).

5. Conclusion

The alignment and translation techniques described in this paper make it possible to generate automatically morphological resources of the biomedical domain for many languages, by exploiting existing multilingual terminologies like the UMLS. Indeed, our approach relies on the use of Kanjis as a pivot language and on alignment algorithm as well as cooccurrence mining. In previous work, we have

Term	DeriF (our translation)	Decomposition in kanjis + generated translation
asbestose	(Part of – Particular type of) chronic condition related to *asbeste*	石綿 'asbestos' 症 'disease'
anurie	Absence of urinary tract	無 'not' 尿 'urine' 症 'disease'
stéréopsie	Condition related to *stère*	立体 'tridimensionnal' 視 'vision'
calcinose	(Part of – Particular type of) chronic condition related to *calcin*	石灰 'calcium' 症 'disease'
pneumopéricarde	(Part of – Particular type of) pericardium related to lung	気 'air' 心膜 'pericardium' 症 'disease'
sarcolemme	(Part of – Particular type of) lemme related to flesh	筋 'muscle' 鞘 'sheath'

Table 3: Examples of analyses produced by DeriF and our resources

shown that the alignment process makes it possible to segment morphologically complex terms and to align the segments to their corresponding Kanjis (Claveau and Kijak, 2011). We have also proved that this way of decomposing the morphological compounds into smaller meaningful units, with Kanjis serving as labeled, could be beneficial to biomedical information retrieval tasks (Claveau, 2012). In this paper, we propose new evaluations of these resources and show that they are also useful for other tasks. In particular, we first come back on the task of linear segmentation of morphological compounds; we show that our automatically built resources, used with simple algorithms, rival the best existing system, without requiring any expert knowledge. We also show that, when combined with two simple analysis rules, the resources make it possible to easily develop very competitive systems to find the hierarchical structure of terms. Last, with simple techniques based on cooccurrence mining in multi-word terms of the UMLS, we build correspondences between morphs and terms (such as 'osis'/disease) which allow to build tools for easily interpret the morphologically complex term for a human reader.
So far, we have generated these resources (probability tables and translation tables) for English, French, Spanish and Portuguese. Their distribution is foreseen but may require the agreement of the copyright holders of the terminologies included in the UMLS MetaThesaurus.

6. References

Baud, R.H., Rassinoux, A.-M., Ruch, P., Lovis, C., and Scherrer, J.-R. (1999). The power and limits of a rule-based morpho-syntactic parser. In *Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association. Transforming Health Care through Informatics. AMIA*, pages 22–26, Washington, DC, USA.

Chiao, Yun-Chuang and Zweigenbaum, Pierre. (2002). Looking for French-English translations in comparable medical corpora. *Journal of the American Medical Informatics Association*, 8(suppl).

Claveau, Vincent and Kijak, Ewa. (2011). Morphological Analysis of Biomedical Terminology with Analogy-Based Alignment. In *Proceedings of RANLP conference*, Hissar, Bulgaria.

Claveau, Vincent and L'Homme, Marie-Claude. (2005). Structuring terminology by analogy-based machine learning. In *Proc. of the 7th International Conference on Terminology and Knowledge Engineering, TKE'05*, Copenhaguen, Denmark.

Claveau, Vincent. (2012). Unsupervised and semi-supervised morphological analysis for Information Retrieval in the biomedical domain. In *Proceedings of the Computational Linguistics (COLING) Conference*, Mumbai, Inde.

Creutz, Mathias and Lagus, Krista. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Technical report, Publications in Computer and Information Science, Report A81, Helsinki University of Technology.

Dal, Georgette and Amiot, Dany. (2008). La composition néoclassique en français et ordre des constituants. In Amiot, Dany, editor, *La composition dans une perspective typologique*, pages 89–113. Arras: Artois Presse Université.

Deléger, Louise, Namer, Fiammetta, and Zweigenbaum, Pierre. (2008). Morphosemantic parsing of medical compound words: Transferring a french analyzer to english. *International Journal of Medical Informatics*, 78:48–55. Supplement 1.

Fradin, Bernard. (2005). On a semantically grounded difference between derivation and compounding. In Dressler, W. U., Katovsky, D., and Rainer, F., editors, *Morphology and its Demarcations*. Amsterdam / Philadelphia: John Benjamins.

Grabar, Natalia and Zweigenbaum, Pierre. (2002). Lexically-based terminology structuring: Some inherent limits. In *Proc. of International Workshop on Computational Terminology, COMPUTERM*, Taipei, Taiwan.

Hathout, Nabil. (2009). Acquisition morphologique à partir d'un dictionnaire informatisé. In *Actes de la conférence Traitement automatique des langues naturelles, TALN'09*, Senlis, France.

Iacobini, Claudio. (1999). Distinguishing derivational prefixes from initial combining forms. In *Proceedings of the First Mediterranean Morphology Meeting*, Mytilene, Greece.

Jiampojamarn, Sittichai, Kondrak, Grzegorz, , and Sherif, Tarek. (2007). Applying many-to-many alignments and hidden Markov models to letter-to-phoneme conversion. In *Proc. of the conference of the North American Chap-*

ter of the Association for Computational Linguistics, Rochester, New York, USA.

Knight, Kevin and Graehl, Jonathan. (1998). Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Kurimo, Mikko, Virpioja, Sami, and Turunen, Ville T. (2010). *(Eds), Proceedings of the MorphoChallenge 2010*. Espoo, Finlande.

Markó, Kornél, Schulz, Stefan, and Han, Udo. (2005). Morphosaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, 44(4).

Morin, Emmanuel and Daille, Béatrice. (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation (LRE)*, 44.

Nakamura-Delloye, Yayoi. (2007). *Alignement automatique de textes parallèles français-japonais*. Thèse de doctorat en linguistique, Université Paris 7.

Namer, Fiammetta. (2007). Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. *Traitement Automatique des Langues – TAL*, 46(2):157–181.

Pearce, Darren. (2002). A Comparative Evaluation of Collocation Extraction Techniques. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 02*, Las Palmas de Gran Canaria, Spain.

Rabiner, Lawrence R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286.

Tsuji, Keita, Daille, Béatrice, and Kageura, Kyo. (2002). Extracting French-Japanese word pairs from bilingual corpora based on transliteration rules. In *Proc. of the 3rd International Conference on Language Resources and Evaluation, LREC'02*, Las Palmas de Gran Canaria, Spain.

Tuttle, Mark, Sherertz, David, Olson, Nels, Erlbaum, Mark, Sperzel, David, Fuller, Lloyd, and Neslon, Stuart. (1990). Using meta-1 – the 1st version of the UMLS metathesaurus. In *Proc. of the 14th annual Symposium on Computer Applications in Medical Care (SCAMC)*, pages 131–135, Washington, USA.